Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE DÉFENSE

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool

Peter Kwantes

Phil Terhaar

Defence R&D Canada
Technical Report
DRDC Toronto TR 2010-188
November 2010

Canada

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool

Peter Kwantes

Phil Terhaar

Defence R&D Canada – Toronto

Technical Report
DRDC Toronto TR 2010-188
November 2010

Principal Author

Original signed by Peter Kwantes

Peter Kwantes

Defence Scientist

Approved by

Original signed by Keith Stewart

Keith Stewart

Head, Socio-Cognitive Systems Section

Approved for release by

Original signed by Dr. Joseph V. Baranski

Dr Joseph V. Baranski
Chair, Knowledge and Information Management Committee
Chief Scientist

Abstract

The *Graphical Overview of the Social and Semantic Interactions of People* (GOSSIP) is a software program designed to help analysts find important entities discussed in a document collection and uncover the nature of the connections among them. It uses a computational model of a semantic system to create "meaning" representations of all the words/terms it encounters in the collection—including proper names. In this report, we demonstrate that the semantic representation of proper names discussed in a document collection can be usefully queried to find out how strongly the entities are associated with a set of user-defined qualities or concepts. We recommend that GOSSIP be trailed in contexts where intelligence analysts or those engaged in influence activities are forced to quickly develop situational awareness about individuals or organizations in a domain from large collections of relevant documents.

Résumé

GOSSIP est un logiciel qui aide les analystes à trouver des entités importantes mentionnées dans un corpus de documents et à découvrir la nature des connexions entre celles-ci. GOSSIP fait appel à un modèle informatique de système sémantique pour représenter la signification, ou le sens, de tous les mots ou expressions qu'il détecte dans le corpus—y compris les noms propres. Le présent rapport a pour but de démontrer qu'il peut être utile d'interroger la représentation sémantique des noms propres mentionnés dans un corpus de documents pour établir le degré de correspondance entre les entités et un ensemble de qualités ou de notions préétablies. Nous recommandons que GOSSIP soit implanté là où les analystes du renseignement ou les personnes engagées dans des activités d'influence doivent élaborer rapidement une connaissance de la situation sur des individus ou des organisations dans un domaine donné à partir d'un corpus importants de documents pertinents.

This page intentionally left blank.

Executive summary

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool:

Peter Kwantes; Phil Terhaar; DRDC Toronto TR 2010-188; Defence R&D Canada – Toronto; November 2010.

Introduction or background: The reports of interest to the intelligence analyst or Influence Activities analysts will generally be analysed to uncover the people and organisations discussed in the report collection, and how these entities are connected to one another. While such information is useful, it does not reveal information about the entities that can be uncovered through automated methods. In this report, we'll demonstrate that an unsupervised model of semantic memory can be used to generate profiles of entities discussed in document collections. Semantic memory refers to memory for the things one knows as opposed to memory for the things one can remember. The past 20 years have seen great advances in our understanding of semantic memory. So much so, that the latest models can create semantic representation for terms in a completely unsupervised fashion. That is, the models can figure out what terms are semantically similar without the model builder hand-wiring any associations among them. GOSSIP is a software tool developed at Defence Research and Development Canada (DRDC) - Toronto that allows the user to see the connections that exist among entities discussed in a large collection of documents. GOSSIP has a model of semantics working in the background processing the documents in the collection. Over the thousands of documents of a collection that are processed, it forms semantic representations for terms and entity names. The semantic representations form a basis upon which to filter documents or entities. In this report, we show that GOSSIP's semantic representations of entities and documents discussed in the text can be queried to find out what concepts connect entities, and more interestingly, it can generate profiles across a set of user-defined qualities. We tested GOSSIP by conducting two empirical studies in which subjects were asked to make judgments about famous names, and about the concepts that connect pairs of famous names.

Results: We found that the information GOSSIP extracted from the document collection was, for the most part, in line with the domain knowledge provided by subjects. In other words, humans and GOSSIP were in close agreement about the material discussed in the document collection.

Significance: We take the results reported here as clear evidence that GOSSIP is a potentially useful tool for quickly establishing situational awareness about people, places and groups discussed in large document collections (situation reports or open source media). Using GOSSIP will help analysts in the Canadian Forces to gain situation awareness about a domain in a timely manner without sacrificing accuracy.

Future plans: It is our goal that the capabilities embodied by GOSSIP will at some point be integrated as a service in existing analysis tools being developed for, and exploited by, the Canadian Forces.

Sommaire

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool:

Peter Kwantes; Phil Terhaar; DRDC Toronto TR 2010-188; R & D pour la défense Canada – Toronto; Novembre 2010.

Introduction ou contexte : En règle générale, l'analyste du renseignement ou celui des activités d'influence examine le corpus de rapports qui l'intéressent dans le but d'y repérer le nom des personnes et des organisations qui y sont mentionnées et de découvrir des « connexions », c'est-à-dire les liens qu'elles présentent entre elles. Bien qu'une telle information soit utile, elle n'apporte rien de plus que ce que des méthodes automatisées permettent d'apprendre à leur sujet. Le présent rapport nous permet de montrer qu'un modèle de mémoire sémantique appliqué sans supervision peut servir à générer les profils des entités mentionnées dans un corpus de documents. Par mémoire sémantique, nous entendons la mémoire qui procède de la connaissance, c'est-à-dire ce que l'on sait, par opposition à celle qui relève des souvenirs, c'est-à-dire ce dont on se souvient. Au cours deux dernières décennies, le savoir humain a réalisé des progrès remarquables dans le domaine de la mémoire sémantique, à telle enseigne que les modèles les plus récents sont en mesure de créer une représentation sémantique des termes en l'absence de toute forme de supervision. Autrement dit, ces modèles découvrent eux-mêmes les termes qui présentent des similitudes sémantiques sans que le concepteur ait à introduire des associations entre ces termes. Mis au point à RDDC Toronto, GOSSIP est un logiciel qui permet de visualiser les connexions entre les entités mentionnées dans un corpus comptant un grand nombre de documents. Son fonctionnement repose sur un modèle sémantique exécuté en arrière-plan qui traite les documents du corpus. Le logiciel crée une représentation sémantique de chaque terme et nom d'entité que contiennent les milliers de documents ainsi traités. De telles représentations sémantiques forment la base à partir de laquelle sont filtrés les documents et entités. Dans le présent rapport, nous montrons que GOSSIP permet d'interroger les représentations graphiques des entités et des documents mentionnés dans un texte et de découvrir ainsi les notions qui lient les entités entre elles. Plus intéressant encore, il permet de générer des profils répondant à un ensemble de caractéristiques préétablies. Nous avons vérifié l'efficacité de GOSSIP sur ces deux plans en menant autant d'études empiriques aux cours desquelles les personnes interrogées devaient exprimer leur opinion sur diverses célébrités et sur la nature des liens qu'évoquent dans leur esprit divers noms de célébrités appariés.

Résultats : Nous avons constaté que les renseignements que GOSSIP extrait du corpus de documents correspondaient dans une large mesure aux connaissances des personnes interrogées dans le domaine. En d'autres mots, ces personnes et GOSSIP étaient presque unanimes quant au contenu mentionné dans le corpus de documents.

Portée : Nous sommes d'avis que les résultats présentés ici démontrent clairement que GOSSIP peut s'avérer utile pour élaborer rapidement une connaissance de la situation sur des gens, des endroits et des groupes mentionnés dans un corpus volumineux de documents, qu'il s'agisse de comptes rendus de situation ou de contenu provenant de sources ouvertes (médias). Grâce à

GOSSIP, les analystes des Forces canadiennes pourront acquérir une connaissance de la situation en temps opportun sans sacrifier l’exactitude des renseignements.

Perspectives : Nous avons pour objectif d’intégrer tôt ou tard sous forme de service les capacités de GOSSIP aux outils d’analyse actuels.

This page intentionally left blank.

Table of contents

List of figures

1 What is semantic memory?

Semantic memory is a specialized memory system that stores, among other things, "meaning" information about the language a person knows. In effect, semantic memory is memory for the things one knows as opposed to memory for the things one can remember. Over the past 20 years we have seen great advances in our understanding of semantic memory. Some of these advances can be attributed to new computational techniques for understanding how people generate meaning representations for words from their exposure to them in language. There are about a dozen computational models of semantic memory. Although the algorithms differ somewhat, all of the models work on the same basic principle: that the contexts in which words are used define their meaning. The new models of semantic memory differ fundamentally from the traditional computational models of semantic memory. In the classic models, semantic memory was treated as a network of nodes, each of which represented some concept. Concepts were connected to one another with a strength that varied as a function of how closely they were associated. The difficulty with the classic models is that they required the model builder to wire them by hand. That is, the connection between the terms *dog* and *cat* had to be established by the model builder. The new models discover such associations automatically. It is the unsupervised nature of these models that makes them valuable as tools in software designed to help understand the contents of large document collections.

Of the models that exist, we have chosen one called, Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) to serve as a semantic system that forms semantic representations for words discussed in reports (e.g., Intelligence reports, open source media). Latent semantic analysis, or LSA, starts by creating a term-by-document matrix from the terms that occur in tens of thousands of documents. Each cell of the matrix contains the frequency with which a particular term occurs in a particular document. The row of cells for a term contains the frequency with which a term occurs across the tens of thousands of contexts or documents under examination.

This term by document matrix is then submitted to a statistical technique called singular value decomposition or SVD. SVD decomposes a term-by-document matrix into three matrices. Without going into too much detail, one of the matrices is a diagonal matrix containing singular values or eigenvalues. There is one for each document because, to differentiate all of the term vectors from each other, there needs to be as many dimensions as there are documents. The eigenvalues vary in value; the higher the value, the more variance in the original matrix is accounted for by that dimension.

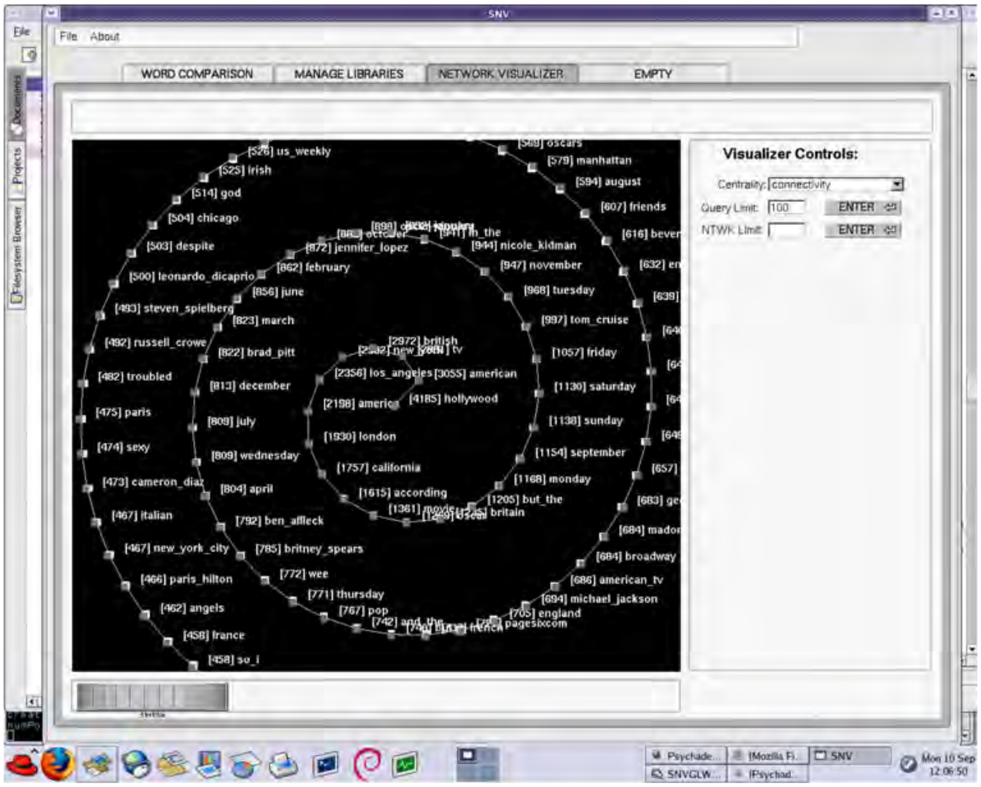
When the three matrices are multiplied together the original matrix is recreated. LSA works by recreating the original matrix on the basis of the top 200 to 300 singular values rather than using all of them. Doing so forces the system to recreate the original matrix on the basis of incomplete information. The resultant matrix is therefore an approximation to the original. The exclusion of so many singular values means that the system must, in a sense, guess as to what the frequencies must be in all the contexts. When it does so, words that the system deduces <u>should</u> occur together in documents or contexts end up having vectors that resemble each other much more than words that do or should not occur together.

LSA is a technique for deriving semantic representations for content words. Typically, the user will apply a list of stop words to exclude function words from the SVD. Function words are words that carry little to no semantic content because they either occur so often or for the purpose of injecting emphasis in the text (e.g., curse words). Words like *a, the, and, he, she,* and *because* are so ubiquitous across contexts, contextual use cannot be used as a basis on which to derive meaning. Likewise, expletives are used to express emphasis and carry little meaning. Proper names are also not given special treatment in models like LSA. In this report we will demonstrate that when treated like terms, proper names contain information that can be queried by the user to extract information about entities that would otherwise only be discovered by reading the documents.

LSA generates semantic representations for terms that take the form of a vector. One can measure how semantically similar two terms are by measuring the similarity of the vectors representing them. Similarity is generally measured using the vector cosine which behaves much like a Pearson correlation coefficient. A cosine of zero, for example, means that there is no similarity between the vectors for two words, and a cosine of 1.0 means that the vectors are perfectly aligned. One can also measure the similarity between a term and a document, or a pair of documents. A document vector is created by summing the vectors for the content words contained in a document. By creating a document vector, one can measure its similarity to a term or concept (which we represent as a summed collection of term vectors) to determine how strongly the ideas expressed by the term/concept are present in the document.

2 What is GOSSIP?

GOSSIP is a software tool developed at DRDC Toronto that allows the user to see the connections that exist among entities discussed in a large collection of documents. A detailed description of GOSSIP can be found in Kwantes (2009, DRDC TR-2009-153). GOSSIP uses a simple visualization interface to show the user what the important entities (i.e., people, places, organizations, or anything with a name), are in a collection of documents, who they are connected to, and the nature of the connections. *Figure 1* shows a screen shot from GOSSIP. The entity at the top of the spiral is the most important entity in the collection of documents. An entity's importance in GOSSIP refers to the number of entities in the collection he or she appears with in the collections of documents.



Centrality: connectivity
Query Limit: 100
NTWK Limit:

Figure1. Screen Shot of GOSSIP

3 What's in a name?

We mentioned above that LSA creates a term-by-document matrix. What is a term? For LSA, a term is any string of alphabetic characters. LSA does not care what language the letter string is written in. All it cares about is that the same letter string is used in the several different contexts/documents being examined. In our treatment of LSA, we applied an algorithm to a document collection that identified entities in the collection and created an entry for them in the term-by-document matrix. Now, when we apply singular value decomposition, proper names will have semantic representations. The question we can ask then is what information is contained in the vector for proper names? The semantic representation of a proper name is constructed no differently from any other word. That is, its contextual use determines semantic representation. Hence, the kinds of words that are used in documents that discuss individuals will play a part in creating the semantic representation.

In what follows, we will demonstrate that models like LSA can be used to provide a shortcut to finding out information about individuals or groups discussed in reports.

3.1 Study 1: What best describes the connections between entities?

One potential use of LSA in GOSSIP as an intelligence analysis tool is to use its semantic representations to uncover information about entities. One aspect that we may wish to know about entities discussed in our collection is: what concept best describes the relationship between a pair of entities? An important related question is: if we were to use LSA as a device to answer such a question, to what extent can we trust the answer it provides? To answer this question, we ran a study in which subjects were asked to provide judgments about the relationship between pairs of Hollywood celebrities. We then trained LSA/GOSSIP on seven years worth of celebrity gossip from the Internet movie database and asked GOSSIP to make the same judgment.

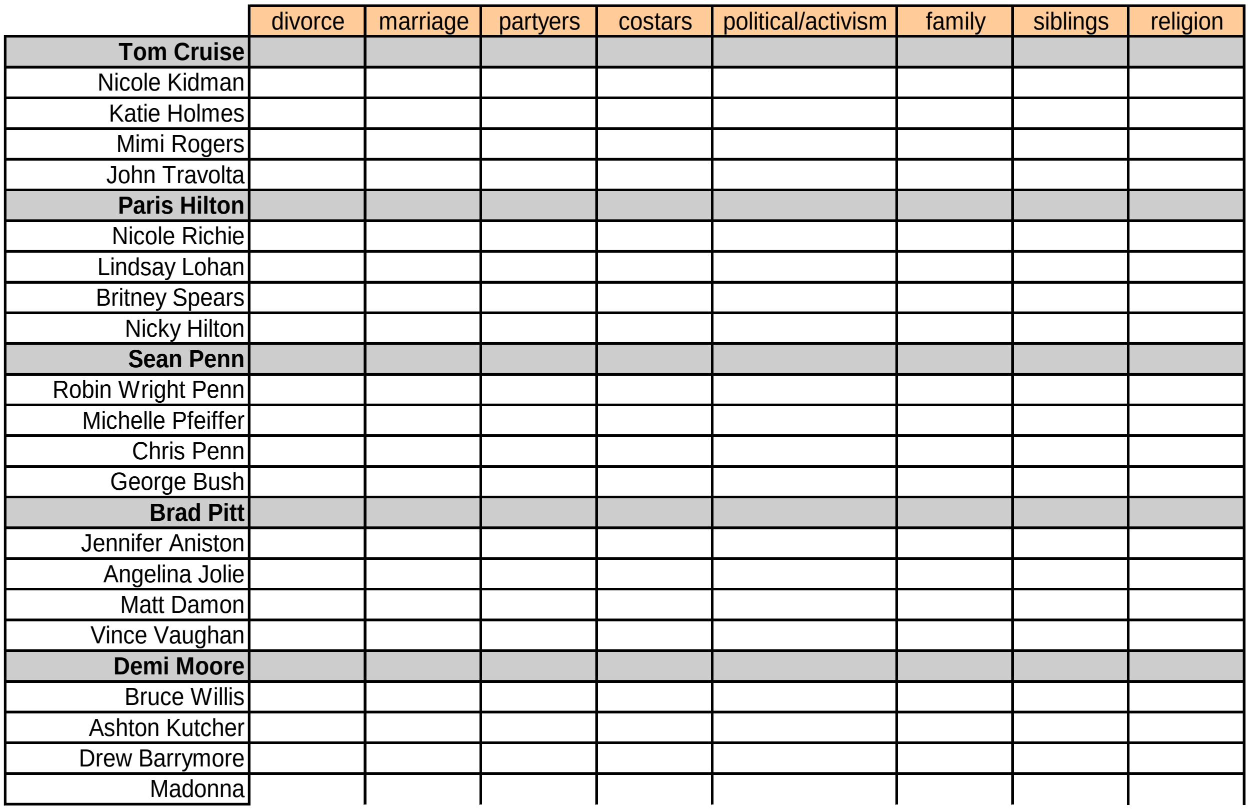
3.1.1 Method

Subjects

Ten employees of DRDC Toronto participated in the study. The study was approved by the DRDC Human Research Ethics Committee. All subjects reported having high familiarity with the domain of celebrity gossip.

Materials.

Stimuli were generated by selecting five celebrity names to serve as stem names. For each stem name, four known associates were selected from the Internet Movie Database (IMDB) news corpus. The eight concepts were also selected, each of which represented a form of relationship (e.g., co-stars, family, etc.) The full list of celebrity names along with the selected concepts are shown in *Figure 2*. This is also the table that was given to subjects in a spreadsheet



	divorce	marriage	partyers	costars	political/activism	family	siblings	religion
Tom Cruise								
Nicole Kidman								
Katie Holmes								
Mimi Rogers								
John Travolta								
Paris Hilton								
Nicole Richie								
Lindsay Lohan								
Britney Spears								
Nicky Hilton								
Sean Penn								
Robin Wright Penn								
Michelle Pfeiffer								
Chris Penn								
George Bush								
Brad Pitt								
Jennifer Aniston								
Angelina Jolie								
Matt Damon								
Vince Vaughan								
Demi Moore								
Bruce Willis								
Ashton Kutcher								
Drew Barrymore								
Madonna								

Figure 2. Table shown to subjects in Study 1.

Procedure

Subject data collection. For each stem name, subjects were asked to select the box indicating which concept best describes his/her relationship with each of the associates. So, for example, the concept of *religion* might be the best concept describing the relationship between stem name, Tom Cruise and his associate, John Travolta because they are both Scientologists.

Model Data. For each stem name, we constructed four LSA vectors: one for each associate. Each vector was created by summing all the LSA vectors for all the terms contained in all the documents that mention the stem name and associate together. Put another way, for each name pair, we created a vector that contained the terms from all the documents in which they were discussed together. The next step was to create vectors for each of the concepts that associated the pairs. First, we created a list of terms for each concept that described the concept. For example, the concept of siblings could be described by the list containing the words, *sibling, brother, sister, stepbrother,* and *stepsister*. We then summed the LSA vectors for the component words to create a vector for the concept. As a final step, we compared the document vector for each name pair to the vectors for each concept by calculating their vector cosine. The vector with the highest cosine across the concepts was taken as the "best" descriptor of the relationship between the associates.

3.1.2 Results and Discussion

The question we want to ask of our data is two-fold: first, to what extent do subjects agree on the concept that best describes the relationships between the names contained in the table? The second question, which follows from the first is: does LSA, as it is embodied in GOSSIP, have at least as much agreement with subjects as subjects have with each other? To answer the first question, we measured the extent to which subjects agreed on the dominant concept. For each possible pair of subjects, we calculated the proportion of characteristics/concepts that were agreed upon as most descriptive of each celebrity name pair in the matrix illustrated in *Figure 1*. Across subjects, there was, on average, 70% agreement about which concept best described the relationship among the celebrity pairs.

For GOSSIP/LSA, we measured similarity between each concept and the 20 document collections that represent all of the text in the collection that discusses each celebrity pair together. As mentioned above, for each celebrity pair, we extracted all of the documents in the collection in which they were mentioned together. The content words' vectors of those documents were then summed to create a vector describing the semantic content of the entire discussion of the pair. Similarity of the pairs' vectors across concepts was measured by calculating the cosine between each pair's vector and the vectors representing each concept. The concept with which the pair vector had the highest cosine was taken as the dominant concept describing the relationship between members of a celebrity pair. To answer the question of how well GOSSIP agrees with subjects, we measured the extent to which the dominant concepts decided upon by GOSSIP were matched to the dominant concepts chosen by subjects. When we calculated the proportion of matches among the concepts, the agreement between GOSSIP and the subjects as a group was also 70%. In other words, GOSSIP's assessment of what concepts best describe the relationship between celebrity name pairs was as consistent with our subjects as our subjects were with each other. This is important, because the strictest test of how well GOSSIP can make a judgment about the nature of relationships among test pairs will take the degree of inter-subject agreement as the acceptable minimal amount of agreement that a useful tool must have.

3.2 Study 2: Profiling of entities from reports with GOSSIP: What’s in a name?

LSA and other such models generate semantic representations for the words they encounter in text. Among the words for which the models will generate semantic vectors, if they are allowed to remain in a document collection, are proper names. The semantic content of a word’s vector is shaped by the other words that appear with it across the thousands of documents in a document collection, and the kinds documents in which it tends to appear.

To evaluate GOSSIP’s ability to generate profiles from intelligence, we made an adjustment to the LSA component. Specifically, we pre-processed the IMDB report corpus to identify proper names. Proper names were transformed into a single term by adjoining the first and last names with an underscore character. For example, Tom Cruise was transformed into TOM_CRUISE. Further, within a given document, every instance of "Tom" or "Mr. Cruise" was also re-written as TOM_CRUISE. After the pre-processing was complete, the algorithm was run to create semantic vectors for all of the terms in the corpus.

The question to ask now is: what information lies in the vector representing a person's name? That is, what information can we learn about the entities being discussed in our reports? In models like LSA, contextual use determines the semantic representations; hence, the semantic vector representing entities' names will be built from a consideration of what kinds of documents discuss them and what kinds of words tend to occur with them across documents.

To test how well GOSSIP/LSA could capture important information in an entity's vector, we measure its similarity to the vectors for several concepts. The set of cosines we get from the comparisons is a profile of the entity. How well the profile describes the entity is determined by having human raters create profiles for the same entity across the same concepts, and seeing how well they agree.

3.2.1 Method

Subjects

The same 10 employees of DRDC Toronto who took part in Study 1 took part in the second study.

Materials

Twenty-four famous names (e.g., Tom Cruise, Brad Pitt, and Angelina Jolie) were selected from the same 7-year collection of gossip articles from the IMDB. We also selected 10 concepts that could be associated with individuals in the Hollywood/Entertainment domain. They were: Action Hero, Beauty, Criminal, Funny, Intelligent, Party Animal, Philanthropy/Activism, Religion, Substance Abuse, and Musical.

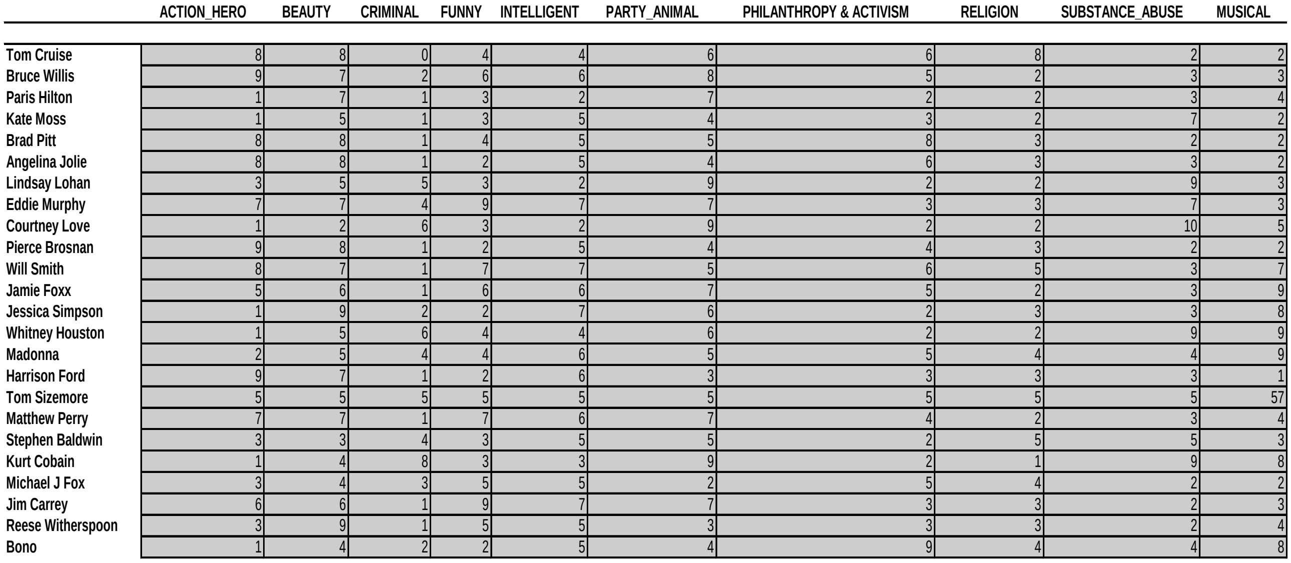
Procedure

Subject Data Collection. Subjects were given a matrix of 24 rows of famous names and 10 columns of qualities (see *Figure 3*). For each famous name, they were asked to rate, on an 11-point scale, how strongly associated the name was with each of the qualities listed in the columns. A 0 on the scale denoted a weak or non-existent association with the quality, and a 10 denoted a strong association. The instructions also made it clear that subjects' ratings were not meant to reflect their opinions about the famous people named in the study. The ratings were strictly meant to reflect the strength of the association between the qualities and the famous names. So, for example, subjects had to rate how strongly someone like Angelina Jolie is associated with the quality of "beauty", not how beautiful they considered her to be.

The ratings from all subjects were averaged to create the "average profile" for each of the famous names. As a supplementary task, subjects were also asked to rate, on two 7-point scales, how familiar each famous name was to them, and how confident they were in the ratings they made for each name.

Model Data

A concept vector was created for each quality by selecting a set of terms that broadly describe the concept, and summing their respective semantic vectors to create a concept vector for each quality. For each famous name, we calculated the vector cosine between the entity's name and the vectors for each quality to create a profile.



	ACTION_HERO	BEAUTY	CRIMINAL	FUNNY	INTELLIGENT	PARTY_ANIMAL	PHILANTHROPY & ACTIVISM	RELIGION	SUBSTANCE_ABUSE	MUSICAL
Tom Cruise	8	8	0	4	4	6	6	8	2	2
Bruce Willis	9	7	2	6	6	8	5	2	3	3
Paris Hilton	1	7	1	3	2	7	2	2	3	4
Kate Moss	1	5	1	3	5	4	3	2	7	2
Brad Pitt	8	8	1	4	5	5	8	3	2	2
Angelina Jolie	8	8	1	2	5	4	6	3	3	2
Lindsay Lohan	3	5	5	3	2	9	2	2	9	3
Eddie Murphy	7	7	4	9	7	7	3	3	7	3
Courtney Love	1	2	6	3	2	9	2	2	10	5
Pierce Brosnan	9	8	1	2	5	4	4	3	2	2
Will Smith	8	7	1	7	7	5	6	5	3	7
Jamie Foxx	5	6	1	6	6	7	5	2	3	9
Jessica Simpson	1	9	2	2	7	6	2	3	3	8
Whitney Houston	1	5	6	4	4	6	2	2	9	9
Madonna	2	5	4	4	6	5	5	4	4	9
Harrison Ford	9	7	1	2	6	3	3	3	3	1
Tom Sizemore	5	5	5	5	5	5	5	5	5	57
Matthew Perry	7	7	1	7	6	7	4	2	3	4
Stephen Baldwin	3	3	4	3	5	5	2	5	5	3
Kurt Cobain	1	4	8	3	3	9	2	1	9	8
Michael J Fox	3	4	3	5	5	2	5	4	2	2
Jim Carrey	6	6	1	9	7	7	3	3	2	3
Reese Witherspoon	3	9	1	5	5	3	3	3	2	4
Bono	1	4	2	2	5	4	9	4	4	8

Figure 3. The matrix of famous names and qualities presented to subjects. The numbers in the matrix in the figure represent the ratings given by one subject for the famous names.

3.2.2 Results and Discussion

How well do subjects and GOSSIP agree on the weights the famous names have on the qualities we chose? The two sources of data are on different scales: integers from 0 to 10 for subjects, and real numbers from 0.0 to 1.0 for the model. As a first step, both sets of data were transformed into *z*-scores. We then expressed each rating for a name as a standard deviation from the average of all the ratings for the name. Done this way, both sources of data are on the same scale. One way we could have standardized the ratings is by calculating *z*-scores from the grand mean of the ratings across qualities and subjects. The problem with doing so is that the magnitudes of the cosines derived from LSA are sensitive to the frequencies of the terms in the document collection, which will skew the ratings. So, for example, the cosine of the vectors for *Religion* and Tom Cruise may be higher than the cosine for *Religion* and Madonna. The difference may occur because one is more strongly associated with *Religion* than the other. Alternatively, it might reflect a difference in the number of qualities with which Madonna and Tom Cruise have strong associations. A term's vector is made up of all the available contextual information taken from the document collection. The information from the different sources is summed in the vector and cannot be separated. As a result, even if Tom Cruise has a stronger association to *Religion* than Madonna, the similarity between Tom Cruise and *Religion* may be lower than the similarity between Madonna and *Religion* if his vector contains more information about other qualities than hers.

We have plotted the profiles from GOSSIP/LSA and subjects as radar plots in *Figure 4*. There is a separate plot for each famous name. Each of the ten axes of a plot represents a quality on which the name was rated. A point close to the graph's origin indicates a low rating on a quality and far points indicate high ratings. When one connects the dots round the axes, it forms a shape that represents the profile of the name across the ten qualities. There are two profiles in each graph, the profile formed by joining open circles are the profiles created by averaging the ratings given by subjects. The closed circles are the profiles provided by GOSSIP/LSA. As is clear in the figure, there is very strong agreement between our subjects and the model for several of the famous names. On average, the correlation, as measured by the Pearson Product Moment correlation, between the values in the profiles for each name against the model was modest, with an average of .67. The values across famous names ranged however from very high, .94, to negative, -.17. In the section that follows, we explore some of the reasons for why subjects and the model disagreed on some qualities of famous people.

Where does the disagreement come from? In order to be useful as an intelligence tool, there needs to be good understanding of where the disagreement between the sources of the ratings comes from. Without it, an analyst could not trust the output of a tool that generates profiles from reports, and it would be useless. In our examination of the data, we have noted seven aspects of the material that would have contributed to a disagreement between human and model.

Conceptual associations: There are some qualities that subjects seem to automatically associate despite being discussed separately in the documents. The prime example in this corpus are the qualities, "Substance Abuse" and "Party Animal". Subjects were overwhelmingly biased to believe that a person who was known to be a substance abuser was also a party animal, or vice versa. The reports, however, did not necessarily make the same association. For example, gossip around Paris Hilton is strongly associated with her partying behaviour. However, the documents that discuss Paris Hilton do not generally include discussions about drug use. As a result, the model does not associate her with drug use. Lindsay Lohan, however, is another story—her drug use and party behaviour go hand-in-hand in the gossip articles, and it shows in her profile.

Double Lives of Entities: Celebrity gossip is not a perfect equivalent for intelligence. GOSSIP/LSA cannot distinguish between an actor and the character he or she plays. For example, in our collection, Tom Hanks may be strongly associated with the quality of Religion. The strong association occurs, not because of Tom Hanks' religion, but because of his role in the religiously themed file, *The DaVinci Code*.

Inadequate concept construction: The concepts we used for this analysis were selected by the experimenters. As such, the degree of association between an entity and a concept/quality will be greatly determined by the quality of the concept. That is, the agreement will depend on the terms the user selects to describe a concept. We recommend that, when using GOSSIP, analysts confer with others to come up with the best list of descriptor words they can.

People don't always follow instructions: We were clear in our instructions to subjects: rate the famous names for how strongly they are associated with each quality. Despite being told not to, however, they might have rated names on the basis of their opinions, not known associations.

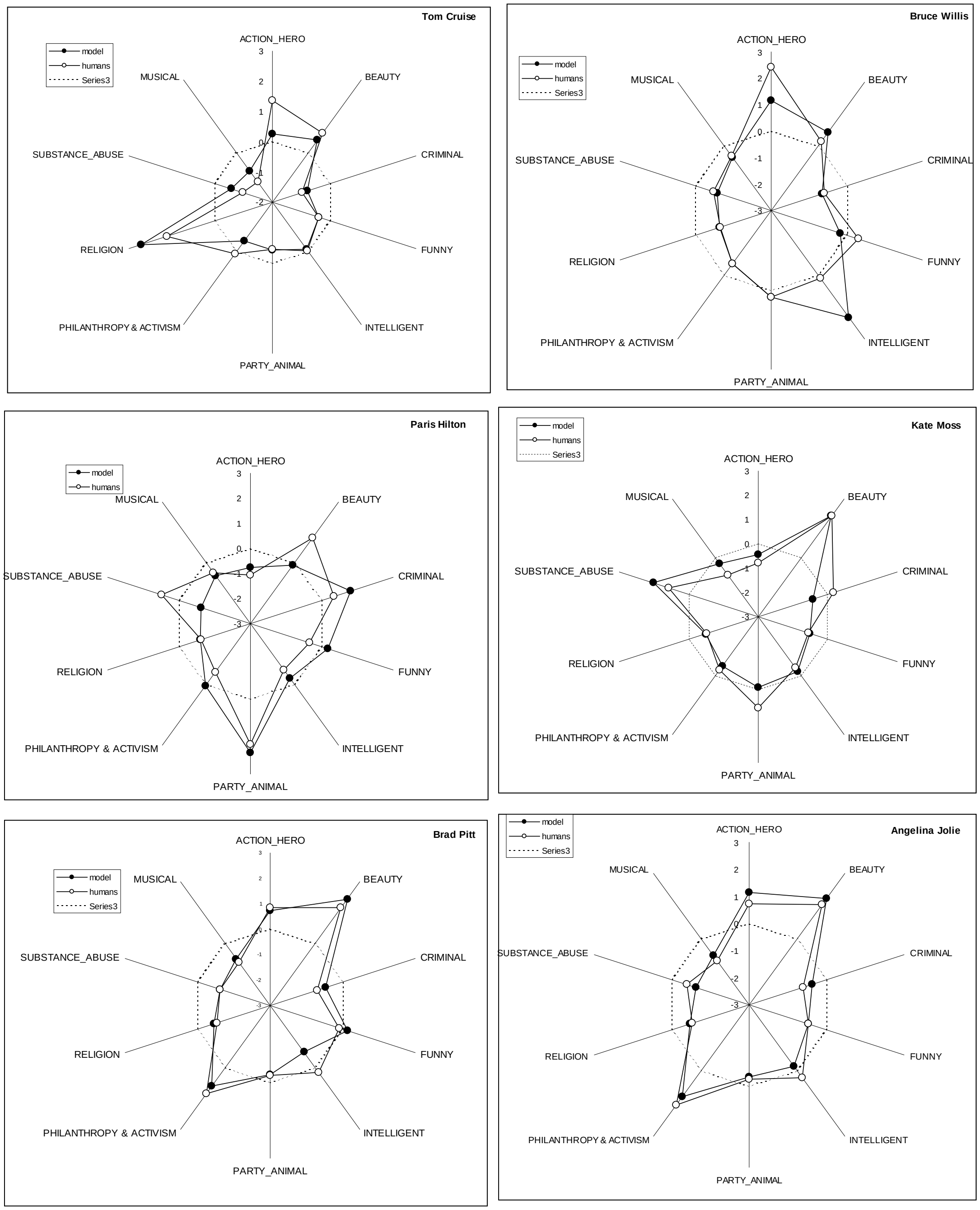


Figure 4. Profiles for famous names generated by subjects and the model. The dashed line circling the graph indicates a z-score of zero on each quality.

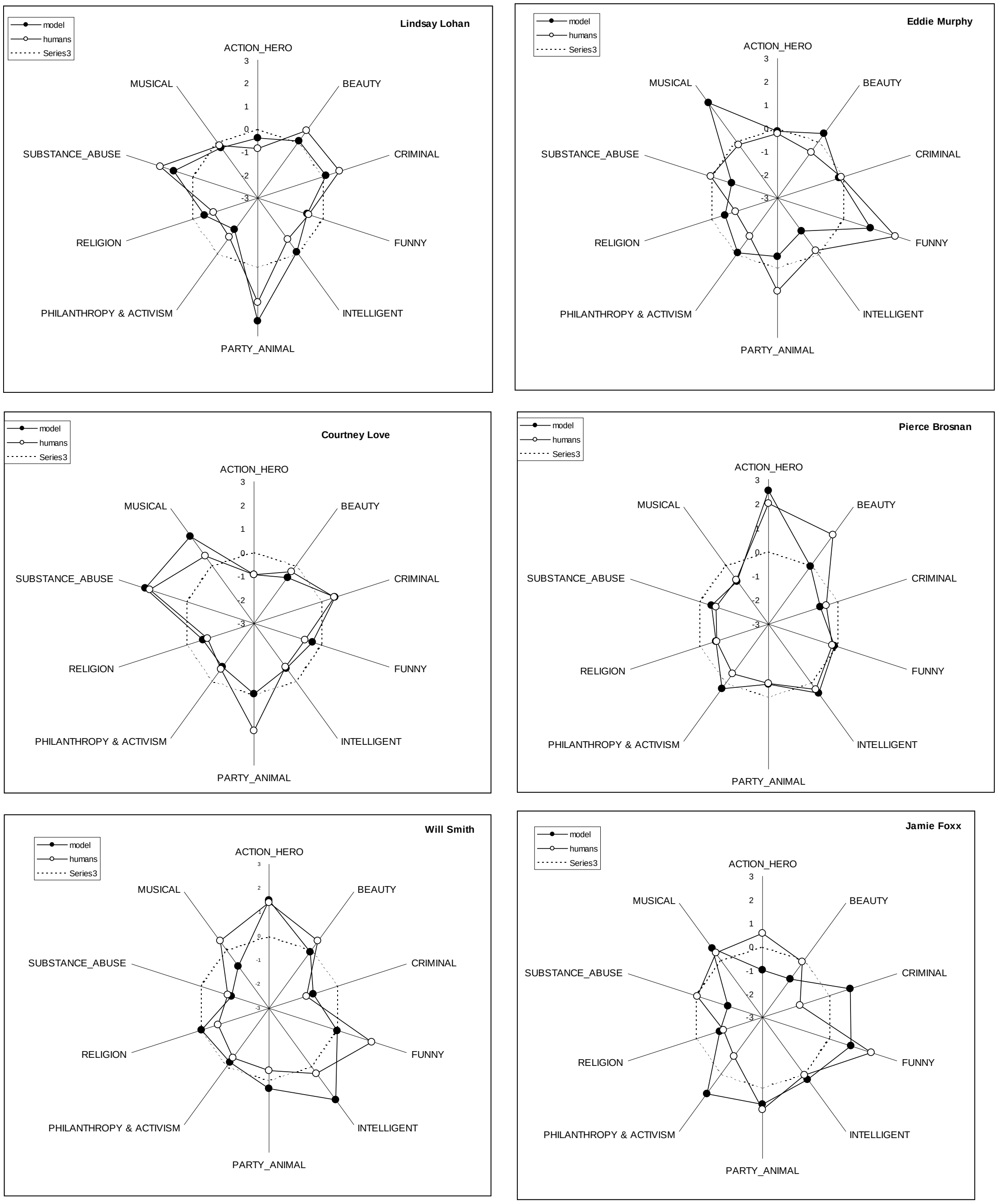


Figure 4. continued. Profiles for famous names generated by subjects and the model. The dashed line circling the graph indicates a z-score of zero on each quality.

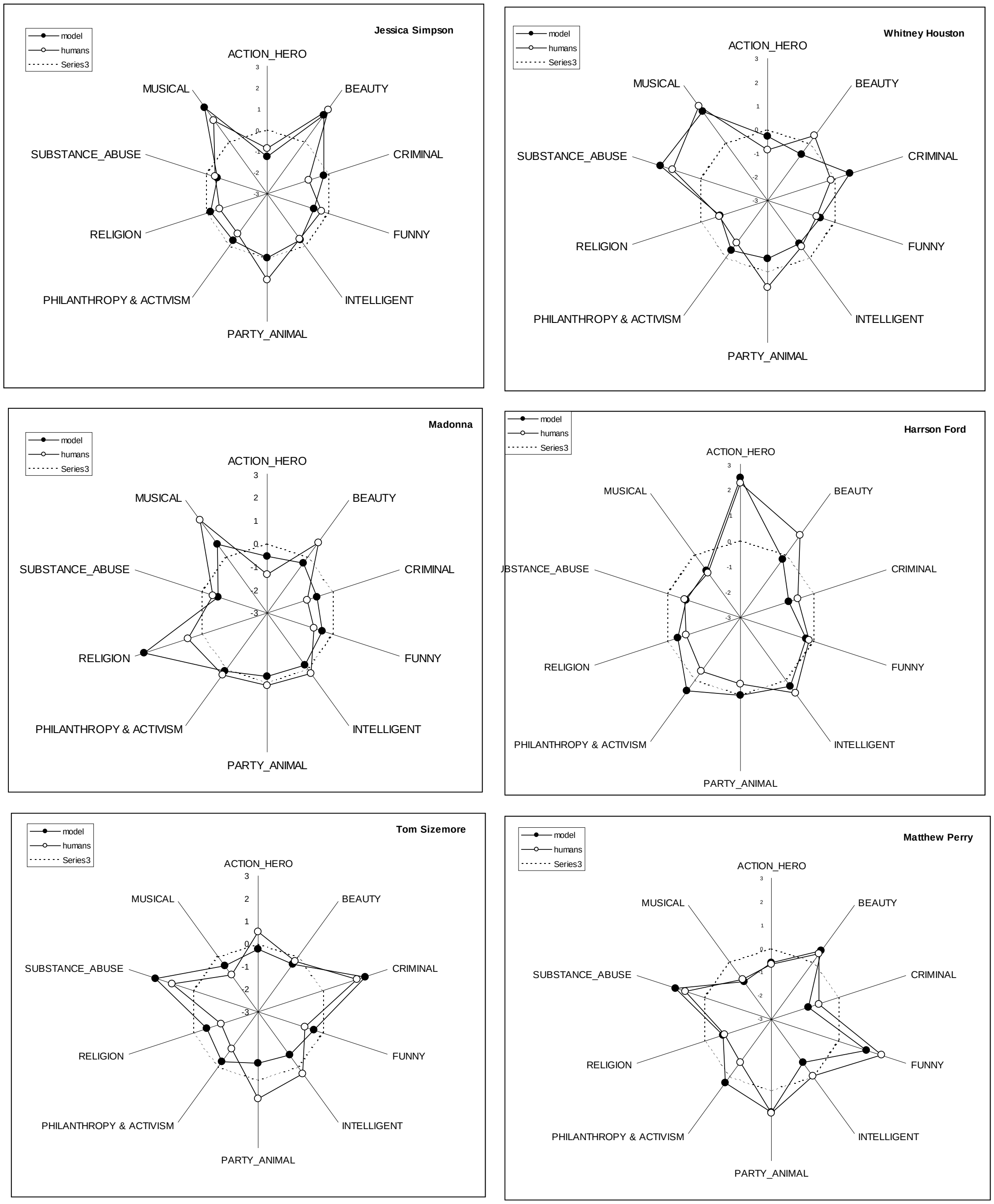


Figure 4. continued. Profiles for famous names generated by subjects and the model. The dashed line circling the graph indicates a z-score of zero on each quality.

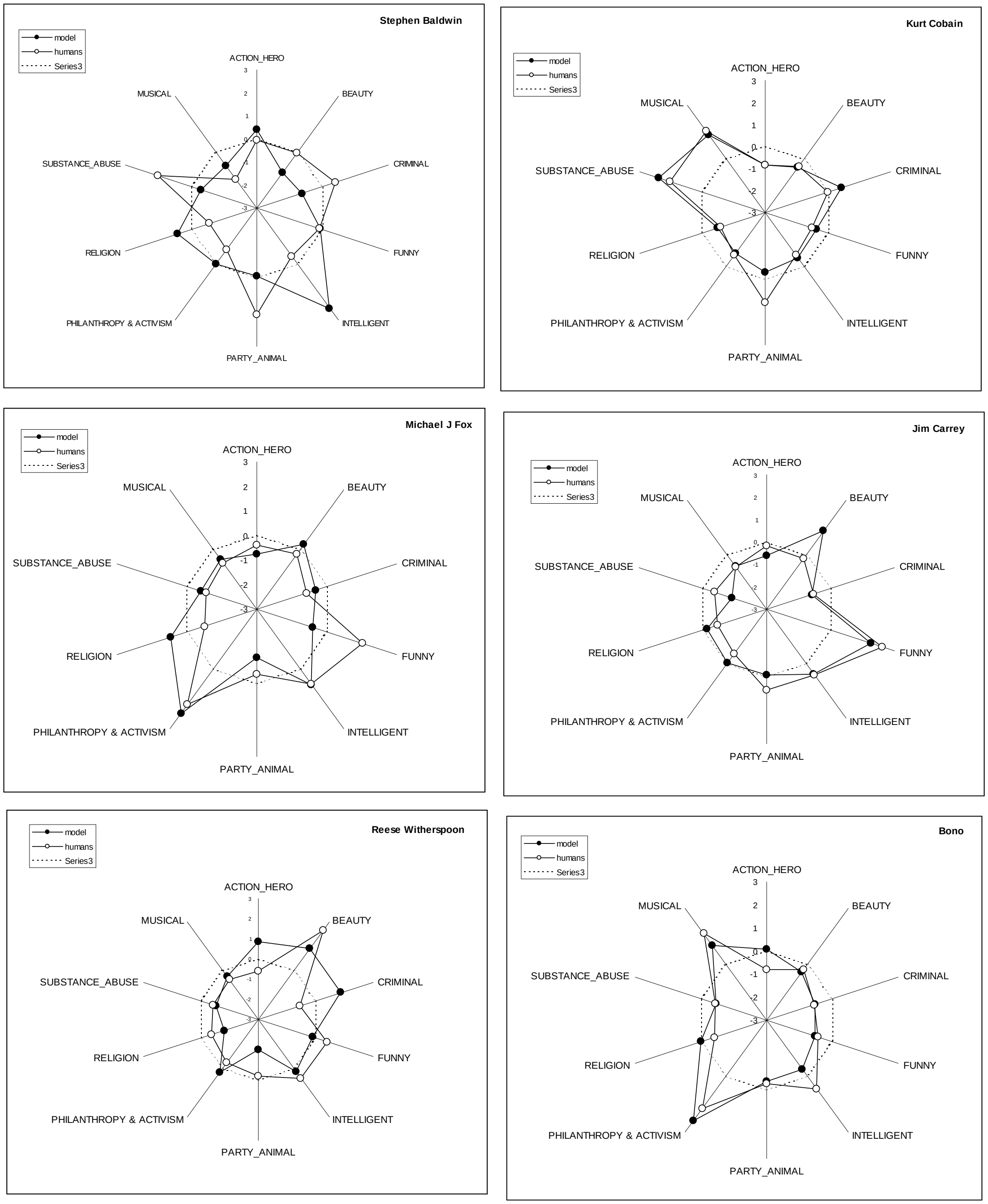


Figure 4. continued. Profiles for famous names generated by subjects and the model. The dashed line circling the graph indicates a z-score of zero on each quality.

So, for example, when rating someone like Angelina Jolie on the quality of *Beauty*, subjects could have made their rating on the basis of how attractive they considered her to be, not on how associated she is with the concept.

Some entities aren't particularly well-known for any of the attributes we selected: It is a fact that some of the famous names we selected for this study are not particularly well-known for any of the qualities we included. For example, the actor Reece Witherspoon is well known, but not for any of the qualities we had included. This was an oversight on our part. When this happens, there is little basis on which the model and human data can agree.

The corpus might have limited documentation for a person: The more a term or entity occurs in a corpus, the more contextual information LSA has to generate a semantic representation. There was a small but reliable tendency for agreement to increase with entity frequency (r = .27, p < .01). In the context of this study, it means that the semantic representations for entities who are discussed often, like Brad Pitt who is mentioned over 800 times, will have vectors containing a rich semantic representation, relative to those for an entity like Stephen Baldwin, who only occurs 8 times, and as a result has very little contextual information upon which to construct a semantic representation. As a result, the disagreement between model and human can come from the poor quality of semantic representation resulting from an entity's infrequent presence in documents.

Familiarity: Perhaps even more important for the agreement between the model and humans is the entities' familiarity to the raters. Although it is a measure correlated with the entity's frequency of occurrence, it taps a slightly different dimension of a famous name. One can be just as familiar with entities that occur half as often as others in the collection. Familiarity was a much better predictor of agreement, r = .41, p < .01, than the frequency of occurrence discussed under the previous point above. *Figure 5* plots the agreement (in the form of a correlation) between GOSSIP and subjects as a function of how familiar entities were judged to be by the subjects. As is clear in the figure, agreement between data and the model increases with the familiarity of the entities.

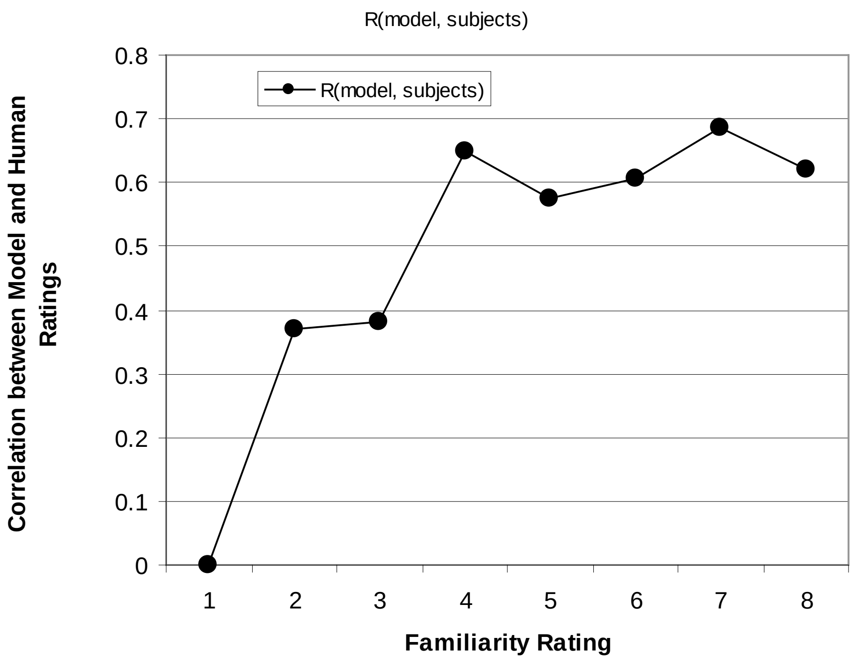


Figure 5. Average agreement between the model and subjects as a function of the familiarity of the entities.

4 General discussion

So what? A skeptical reader may, at this point, be thinking: the only thing these data show is that DRDC employees read the Hollywood gossip reports. Why should I be impressed by these data? The intention behind using a tool such as this one is not to confirm what one already knows. Instead, it is to discover what one does not know. In other words, the usefulness of this aspect of GOSSIP will be clear in situations where the analyst knows little about the entities being discussed in text, and wants a rapid means by which to generate an impression of the qualities associated with them.

We would argue that GOSSIP provides an accurate profile of entities discussed in a corpus. To be believable, however, we have learned from this exercise that the number of times an entity is discussed in the document collection needs to be taken into consideration when making a judgment about its accuracy. The fewer times an entity is mentioned, the less accurate will be the profile. It is therefore important that the analyst be able to explore the documents that mention an entity to verify or clarify the qualities that GOSSIP associates with the person. From the analysts perspective however, this is not a serious setback because GOSSIP provides the user with an easy way of exploring the documents through the visualization interface. As our techniques for analyzing the semantic content of documents becomes more sophisticated, we can foresee being able to provide analysts with a means of assigning measures of confidence to the profiles generated by GOSSIP.

One interesting point to mention about the profiles GOSSIP/LSA generates is that their veracity is relative. That is, the profile of the same individual may be very different across the same qualities depending on the document collection that is being used to generate them. From an analysis perspective, it may be useful to examine how the same entities are treated in different source documents; currently there are no tools other than GOSSIP that could provide such information.

5 Conclusions and recommendations

GOSSIP is a Linux-based software program designed to help analysts find important entities discussed in a document collection and uncover the nature of the connections among them. It uses LSA as a semantic system to create “meaning” representations of all the words/terms and proper names it encounters in the collection. We recommend applying the program for trial in operational contexts to establish its usefulness. In particular, we believe that GOSSIP would be valuable in the analysis of Situation Reports, from which an analyst might wish to find and examine connections among entities discussed over long spans of time. We also believe that GOSSIP would be useful in the Reading-In process by which the analyst might need to gain a rapid understanding of a domain by reading/processing media articles written for and by those in certain contexts.

References

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Kwantes, P.J. (2009) Close-out Report for TIF 15da05: Automatic extraction and visualization of concepts for the operational commander. DRDC Toronto Technical Report, TR 2009-153.

DOCUMENT CONTROL DATA

(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document, Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's document, or tasking agency, are entered in section 8.)

Publishing: DRDC Toronto

Performing: DRDC Toronto

Monitoring:

Contracting:

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification is indicated by the appropriate abbreviation (S, C, R, or U) in parenthesis at the end of the title)

Entity Profiling for Intelligence Using the Graphical Overview of Social and Semantic Interactions of People (GOSSIP) Software Tool (U)
(U)

4. AUTHORS (First name, middle initial and last name. If military, show rank, e.g. Maj. John E. Doe.)

Peter Kwantes; Phil Terhaar

5. DATE OF PUBLICATION (Month and year of publication of document.)

November 2010

6a NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

40

6b. NO. OF REFS (Total cited in document.)

3

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of document, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Report

8. SPONSORING ACTIVITY (The names of the department project office or laboratory sponsoring the research and development − include address.)

Sponsoring:

Tasking:

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant under which the document was written. Please specify whether project or grant.)

15ah

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document)

DRDC Toronto TR 2010−188

10b. OTHER DOCUMENT NO(s). (Any other numbers under which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on the dissemination of the document, other than those imposed by security classification.)

Unlimited distribution

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11), However, when further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited announcement

DOCUMENT CONTROL DATA

(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

(U) GOSSIP is a software program designed to help analysts find important entities discussed in a document collection and uncover the nature of the connections among them. It uses computational model of semantic system to create “meaning” representations of all the words/terms it encounters in the collection—including proper names. In this report we demonstrate that the semantic representation of proper names discussed in a document collection can be usefully queried to find out how strongly the entities are associated with a set of user−define qualities or concepts. We recommend that GOSSIP be trailed in contexts where intelligence analysts or those engaged in influence activities are forced to quickly develop situational awareness about individuals or organizations in a domain from large collections of relevant documents.

(U) GOSSIP est un logiciel qui aide les analystes à trouver des entités importantes mentionnées dans un corpus de documents et à découvrir la nature des connexions entre celles−ci. GOSSIP fait appel à un modèle informatique de système sémantique pour représenter la signification, ou le sens, de tous les mots ou expressions qu’il détecte dans le corpus—y compris les noms propres. Le présent rapport a pour but de démontrer qu’il peut être utile d’interroger la représentation sémantique des noms propres mentionnés dans un corpus de documents pour établir le degré de correspondance entre les entités et un ensemble de qualités ou de notions préétablies. Nous recommandons que GOSSIP soit implanté là où les analystes du renseignement ou les personnes engagées dans des activités d’influence doivent élaborer rapidement une connaissance de la situation sur des individus ou des organisations dans un domaine donné à partir d’un corpus importants de documents pertinents.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

(U) Profiling, intelligence, semantics, target audience analysis